Dorsolateral and ventromedial prefrontal cortex orchestrate normative choice

Baumgartner, Thomas ; Knoch, Daria ; Hotz, Philine ; Eisenegger, Christoph ; Fehr, Ernst

Abstract: Humans are noted for their capacity to over-ride self-interest in favor of normatively valued goals. We examined the neural circuitry that is causally involved in normative, fairness-related decisions by generating a temporarily diminished capacity for costly normative behavior, a 'deviant' case, through non-invasive brain stimulation (repetitive transcranial magnetic stimulation) and compared normal subjects' functional magnetic resonance imaging signals with those of the deviant subjects. When fairness and economic self-interest were in conflict, normal subjects (who make costly normative decisions at a much higher frequency) displayed significantly higher activity in, and connectivity between, the right dorsolateral prefrontal cortex (DLPFC) and the posterior ventromedial prefrontal cortex (pVMPFC). In contrast, when there was no conflict between fairness and economic self-interest, both types of subjects displayed identical neural patterns and behaved identically. These findings suggest that a parsimonious prefrontal network, the activation of right DLPFC and pVMPFC, and the connectivity between them, facilitates subjects' willingness to incur the cost of normative decisions.

DOI: https://doi.org/10.1038/nn.2933

The Mentalizing Network Orchestrates the Impact of Parochial Altruism on Social Norm Enforcement

[1,2]Thomas Baumgartner, [2,3]Lorenz Götte,
[2]Rahel Gügler, [2]Ernst Fehr

[1] Department of Psychology, Laboratory for Social and Affective Neuroscience, University of Basel, Switzerland

[2] Department of Economics, Laboratory for Social and Neural Systems Research, University of Zurich, Switzerland

[3] Department of Economics, University of Lausanne, Switzerland

Keywords: Parochial altruism, ingroup favoritism, outgroup hostility, third-party punishment, punishment network, justification, mentalizing network, fMRI, neural circuitry, social neuroscience, neuroeconomics

Corresponding authors

Thomas Baumgartner
Department of Psychology
Laboratory for Social and Affective Neuroscience
University of Basel
Birmannsgasse 8
CH-4055 Basel
Phone: +41 / 61 / 26 70 288
Email: t.baumgartner@unibas.ch

Ernst Fehr
Department of Economics
Laboratory for Social and Neural Systems Research
University of Zürich
Blümlisalpstrasse 10
CH-8006 Zürich
Phone: +41 / 44 / 634 37 09
Email: ernst.fehr@econ.uzh.ch

Abstract

Parochial altruism – a preference for altruistic behavior towards ingroup members and mistrust or hostility towards outgroup members – is a pervasive feature in human society and strongly shapes the enforcement of social norms. Since the uniqueness of human society critically depends on the enforcement of norms, the understanding of the neural circuitry of the impact of parochial altruism on social norm enforcement is key, but unexplored. To fill this gap, we measured brain activity with functional magnetic resonance imaging (fMRI) while subjects had the opportunity to punish ingroup members and outgroup members for violating social norms. Findings revealed that subjects' strong punishment of defecting outgroup members is associated with increased activity in a functionally connected network involved in sanction-related decisions (right orbitofrontal gyrus, right lateral prefrontal cortex, right dorsal caudatus). Moreover, the stronger the connectivity in this network, the more outgroup members are punished. In contrast, the much weaker punishment of ingroup members who committed the very same norm violation is associated with increased activity and connectivity in the mentalizing-network (dorsomedial prefrontal cortex, bilateral temporo-parietal junction), as if subjects tried to understand or justify ingroup members' behavior. Finally, connectivity analyses between the two networks suggest that the mentalizing-network modulates punishment by affecting the activity in the right orbitofrontal gyrus and right lateral prefrontal cortex, notably in the same areas showing enhanced activity and connectivity whenever third-parties strongly punished defecting outgroup members.

Introduction

The uniqueness of human society is critically dependent on the development, compliance, and enforcement of elementary social norms and the associated altruistic behavior (Fehr and Fischbacher 2003; Fehr and Fischbacher 2004a). For instance, humans are willing to punish violators of social norms even at substantial personal costs (Boyd, et al. 2003; Fehr and Gächter 2002; Henrich 2006). A key element of the enforcement of many social norms, such as food-sharing norms in hunter-gatherer societies (Hill 2002; Kaplan, et al. 2000), is that people punish norm violators not only for direct transgressions against the punisher himself (termed second-party punishment), but also for norm violations against others (termed third-party punishment) (Bendor and Swistak 2001; Sober and Wilson 1998). Norm enforcement requires that even third parties – who are neither economically, physically, nor psychologically affected by the violations – be willing to punish (Fehr and Fischbacher 2004b; Henrich, et al. 2006). Thus, third-party punishment greatly enhances the scope for norms that regulate human behavior. In fact, some researchers view the existence of third-party sanctions as the decisive factor for the enforcement of social norms in human society because second-party punishment strategies are not evolutionarily stable, while strategies involving third-party sanctions are stable (Bendor and Swistak 2001).

Experimental evidence from laboratory (Brewer 1979; Chen and Li 2009; Kinzler, et al. 2007; Koopmans and Rebers 2009; Tajfel, et al. 1971; Tajfel and Turner 1979) and field studies (Bernhard, et al. 2006; Fehr, et al. 2008; Goette, et al. 2006) demonstrates that parochial altruism strongly shapes the compliance and enforcement of social norms. Parochial altruism constitutes a persuasive psychological phenomenon which is qualified by a preference for altruistic behavior towards the members of one's ethnic, racial, or any other social group, combined with a tendency for indifference, mistrust, or even hostility toward outgroup members (Brewer 1999; Hewstone, et al. 2002). For example, a recent third-party punishment experiment in Papua New Guinea revealed strong favoritism toward a subject's own linguistic group in giving to others, and significantly greater punishment of individuals from another linguistic group (in comparison to those from the subject's own group) who committed a norm violation toward the subject's ingroup members (Bernhard, et al. 2006). The importance of parochial altruism for the understanding of human society is corroborated by recent theoretical and experimental research that has closely tied outgroup hostility to the evolution of human prosociality within groups (Boyd, et al. 2003; Choi and Bowles 2007) and prosociality within groups (ingroup favoritism) to the evolution of cultural groups (Efferson,

et al. 2008).

In view of the importance of third-party punishment for the enforcement of social norms, its parochial and altruistic nature, and the evidence for the co-evolution of parochialism, altruism, and cultural groups, we conjecture that humans have developed elaborate neural mechanisms for social cognition that modulate third-parties' norm enforcement behavior dependent on the group affiliation of the norm violator and his or her victim. Although recent studies have fundamentally improved our knowledge of how the brain modulates norm compliance (Baumgartner, et al. 2009; Baumgartner, et al. 2008; Delgado, et al. 2005; Harbaugh, et al. 2007; King-Casas, et al. 2005; Rilling, et al. 2002; Spitzer, et al. 2007) and norm enforcement (Buckholtz, et al. 2008; de Quervain, et al. 2004; Fehr and Camerer 2007; Knoch, et al. 2007; Knoch, et al. 2006; Rangel, et al. 2008; Sanfey 2007; Sanfey, et al. 2003; Strobel, et al. 2011) they do not examine the parochial nature of this phenomena. There is also an important literature examining the neural circuitry of the cognitions involved in the evaluation of faces from distinct races (Cunningham, et al. 2004; Golby, et al. 2001; Phelps, et al. 2000), the judgment of people belonging to other races (Eberhardt 2005; Freeman, et al. ; Ito and Bartholow 2009; Lieberman, et al. 2005; Richeson, et al. 2003), prejudice (Beer, et al. 2008) the evaluation of very poor and "disgusting" outgroups such as addicts and beggars in dirty clothes (Harris and Fiske 2006), and the general evaluation of ingroup-outgroup interactions (Mathur, et al. ; Van Bavel, et al. 2008) but none of the individuals in these studies had to make *costly* punishment decisions that involved real costs and benefits for themselves or for others. In these studies there was thus not trade off between the individual punishers' self-interest, which suggests that he should not punish at all, and the punishers' altruistic concerns, which suggest that he should protect the victim of norm violations. It is exactly this willingness to incur the cost of altruistic norm enforcement which renders altruistic third party punishment a powerful evolutionary force ((Bendor and Swistak 2001; Sober and Wilson 1998).

By using a third-party punishment paradigm (Fehr and Fischbacher 2004b) with manipulation of group membership (in-/outgroup) (Bernhard, et al. 2006; Goette, et al. 2006), the involvement of real monetary stakes, and the requirement to curb immediate self-interests in order to enforce a social norm, the present study is the first to uncover the neural circuitry of parochial altruism and its impact on the enforcement of social norms. For that purpose, we exploited the fact that individuals are randomly assigned to real social groups (platoons)

during a four-week phase of officer training in the Swiss Army. During this training course, officer candidates interact almost exclusively with members of their own platoon and social ties within the platoon form very quickly (for details see method section). The applied third-party punishment paradigm consists of *two decision stages* – one conducted during the third or fourth week of the training course and one conducted in the functional magnetic resonance imaging (fMRI) scanner immediately after the end of the training course (within 5 days). All interactions were anonymous and one-shot.

During the *first decision stage*, the officer candidates played a simultaneous Prisoner's Dilemma Game (PDG). Two players A and B were each endowed with 20 points each and had to decide simultaneously whether to keep the points or pass all of them to the other player. Passed points were doubled. Thus, keeping the points equals defection (denoted as D throughout the paper) and passing the points equals cooperation (denoted as C throughout the paper). For example, if A retained the 20 points while B transferred the 20 points (behavioral pattern DC), then A earned a total of 60 points (40 points from the transfer and plus the original 20 points) and B earned nothing. Thus, irrespective of what the other player did, a player in the first-stage was always better off if he kept the endowment for himself, but if both players kept their endowments they only earned 20 points each (behavioral pattern DD), whereas if they both cooperated and transferred their endowments, each earned 40 points (behavioral pattern CC).

During the *second decision stage*, some of the officer candidates (16 subjects) were invited to the fMRI scanner and received the opportunity, in the role of a third-party (player C), to punish player A's or player B's behavior by assigning punishment points. For that purpose, player C received an endowment of 10 points at the beginning of each punishment trial (30 trials in total, in each of which player C faced the previous decisions of different players A and B), which C could use to finance the assignment of punishment points. Assigning 1 punishment point cost player C 1 point and cost the sanctioned player 3 points. Importantly, player C only could punish the behavior of one player (either A or B) during each of the played punishment trials. In order to simplify the nomenclature, we recoded all player C's decisions in such a way that A always refers to the player that C can punish, while B always refers to the player that C cannot punish. Please note that all players (players A, B, and C) were paid according to their decisions and those of their interaction partners. Thus, no deception of the subjects occurred in this study.

As we wanted to examine the neural circuitry of parochial altruism, third-parties in the fMRI scanner were confronted with the following three group constellations: (1) All three players in the game are from the same platoon (group constellation ABC, depicted in green color throughout the paper). (2) Only players A and C are from the same platoon, while player B is an outgroup member (group constellation AC, depicted in blue color throughout the paper). (3) Only players B and C are from the same platoon, while player A is an outgroup member (group constellation BC, depicted in red color throughout the paper). Because player C (the third-party) and player A (who can be punished, see above) are from the same group in the group constellations ABC and AC, we will in the following refer to these two group constellations as in-group constellations. In contrast, we will refer to the group constellation BC as out-group constellation because here player C (the third-party) and player A (who can be punished, see above) are from different groups. Please see Fig. 1 for a summary of the design and an example of a decision screen the third-parties saw during the scanning session.

Based on the discussed behavioral literature (Bernhard, et al. 2006; Fehr and Fischbacher 2004b; Goette, et al. 2006), we expect third-parties to show a parochial punishment pattern, qualified by a particular strong impact of group membership on punishment decisions in the DC condition, when player A defects and player B cooperates. In detail, we hypothesize that outgroup members who defect against cooperating ingroup members are punished much more severely than ingroup members who commit the same norm violation – a hypothesis the behavioral analyses strongly confirm (see results section for details). The key question this study therefore allows to answer is which brain circuits modulate this highly distinctive parochial punishment pattern.

Recent neuroimaging studies on second-party norm enforcement suggest that two brain regions in particular play a decisive and functionally distinctive role in punishment-related decision processes; these are the right lateral PFC (Knoch, et al. 2007; Knoch, et al. 2006; Sanfey, et al. 2003) and the dorsal caudatus (de Quervain, et al. 2004; Seymour, et al. 2007). Evidence from transcranial magnetic stimulation (TMS) (Knoch, et al. 2006; van 't Wout, et al. 2005) and transcranial direct current stimulation (tDCS) (Knoch, et al. 2007) studies suggests that the right lateral PFC is causally involved in (costly) norm enforcement behavior by modulating the weight of self-interest in the decision process. On the other hand, punishment-related activity in the dorsal caudatus [a brain region strongly implicated in the processing of rewards that accrue as a result of goal-directed actions (Fliessbach, et al. 2007;

Kawagoe, et al. 1998; O'Doherty, et al. 2004; Schultz and Romo 1988)] in combination with behavioral and questionnaire measures suggests that people derive satisfaction from punishing norm violators (de Quervain, et al. 2004; Singer, et al. 2006). We expect similar processes and associated brain activity patterns in the right lateral PFC and the dorsal caudatus when third-parties strongly punish outgroup members (for defecting against cooperating ingroup members). Similar to second-parties, third-parties in our paradigm also have to modulate the weight of self-interest in the punishment decision – a decision process which activity in the dorsal caudatus might also reinforce and motivate.

In addition to similar processes assumed to be necessary for the implementation of both second- and third-party punishment behaviors, we expect that the parochial nature of altruistic norm enforcement requires further psychological and cognitive processes and associated brain activity patterns, as the *very same* norm violation incurs much more severe punishment when an outgroup member is the perpetrator. We assume that at least two additional key processes must take place in the third-party brain in order to mediate such a highly distinctive punishment behavior.

First, we expect the very same norm violation to be evaluated differentially dependent on the norm violator's group affiliation, that is more negatively when an outgroup member is the norm violator and/or more positively when an ingroup member commits the very same norm violation. If this hypothesis is correct, we would expect a differential activity pattern in those regions of the brain strongly associated with a decision-relevant reflective evaluation process, including in particular ventral and lateral regions of the orbitofrontal gyrus (OFG) (Hare, et al. 2009; Kringelbach 2005; Liu, et al. 2007; Plassmann, et al. 2007; Rangel, et al. 2008). More precisely, there is some evidence for a medial-lateral distinction in the OFG, such that activity in medial areas is related to positive evaluation processes, while activity in lateral areas is related to negative evaluation processes (Kringelbach 2005; Liu, et al. 2007; Rilling, et al. 2007; Spitzer, et al. 2007). We thus hypothesize that the increased negative evaluation of outgroup members' norm violations might be associated with increased activity in lateral areas of the OFG.

Second, we expect this group-dependent evaluation process and associated highly distinctive punishment pattern to be associated with a mentalizing process that attempts to justify these highly distinctive punishment decisions. In particular, we assume that a process must be

executed in the third-party brain that justifies the very lenient punishment of ingroup member's defective behavior. Such a justification process might include the search for mitigating reasons or personality characteristics that may be able to excuse the defective behavior. Thus, third-parties may try to understand the underlying reasons and intentions of ingroup member's defective behavior and may consequently find a justification for the defective act. If this hypothesis is correct, we would expect to find increased activity in the mentalizing network of the brain, including the dorsomedial PFC (DMPFC) and the bilateral temporo-parietal junction (TPJ), because this kind of psychological processes - inferring temporary goals, intentions, desires as well as more enduring dispositions of others - have been strongly and consistently associated with this brain network (Gallagher and Frith 2003; Van Overwalle 2009). Finally, if the assumed mentalizing/justification process indeed occurs in the third-party brain, resulting in a reduced negative evaluation and associated reduced punishment behavior, we would expect the areas of the mentalizing network to be functionally connected with the punishment-related areas in the ventral and lateral PFC and/or dorsal caudatus. More precisely, if the mentalizing/justification network is indeed recruited to reduce the punishment of ingroup members' defective behavior, we would expect to find evidence that the mentalizing network controls/downregulates the areas involved in punishment-related decision processes.

Taken together, we hypothesize that the parochial nature of altruistic norm enforcement in the third-party brain is orchestrated by functionally connected brain areas involved in modulating the weight of economics self-interests (right LPFC), the motivational aspect of punishment (dorsal caudatus), negative and positive evaluation (OFG) and mentalizing (DMPFC, bilateral TPJ) processes. More precisely, we hypothesize that **(1)** the increased punishment of defecting outgroup members is associated with increased activity and connectivity in brain areas known to play key and functionally distinct roles in punishment-related decision processes (lateral OFG, lateral PFC and caudatus). In sharp contrast, we expect that **(2)** the reduced punishment of defecting ingroup members is associated with increased activity and connectivity in key areas of the mentalizing network (DMPFC, TPJ). Finally, we hypothesize that **(3)** the mentalizing network might accomplish this reduced punishment of ingroup members by modulating/ down-regulating the activity in parts of the punishment network.

Results

Behavioral results: Parochial punishment patterns

Our main behavioral measure consists of each third-party's average punishment points, broken down into the three group constellations (BC, AC, ABC) and four behavior patterns based on player A's and B's decision to cooperate or defect (CC, CD, DC, DD). On the basis of these average punishment points, we calculated a two-way repeated measures of ANOVA with within-subject factor group constellation (BC, AC, ABC) and within-subject factor behavior (CC, CD, DC, DD). Results revealed significant main effects of group constellation ($F_{(2,14)}$ = 5.36, p = 0.019, ETA^2 = 0.43) and behavior ($F_{(3,13)}$ = 10.34, p = 0.001, ETA^2 = 0.71), as well as a significant interaction effect of group constellation × behavior ($F_{(6,10)}$ = 4.20, p = 0.023, ETA^2 = 0.72). As expected, these main and interaction effects demonstrated the following behavioral punishment pattern (see Fig.2). Cooperative behavioral decisions by player A (behavior patterns CC and CD) resulted in no impact of group membership on punishment behavior (paired t-test behavior CC: all p > 0.11; paired t-test behavior CD: all p > 0.09). This was due to the fact that, irrespective of whether player A was an outgroup or ingroup member, he was not punished for a cooperative behavioral decision (all simple t-tests versus 0: p > 0.11). In sharp contrast, when player A defected and Player B cooperated (behavior pattern DC), group membership had a strong impact on punishment decisions. Outgroup members were punished much more severely for defecting than ingroup members (paired t-test: BC vs AC: p < 0.000, ETA^2 = 0.63; BC vs ABC: p < 0.000, ETA^2 = 0.66). This effect of group membership remained present when both players defected (behavioral pattern DD). The magnitude of the group effect was markedly reduced, however (paired t-test: BC vs AC: p = 0.037, ETA^2 = 0.26, BC vs ABC: p = 0.069, ETA^2 = 0.20), indicated by increased p-values and strongly reduced effect size measures (ETA^2). Finally, there was no significant difference in punishment behavior with respect to the two ingroup constellations (paired t-test behavior DC: AC vs ABC: p = 0.64; paired t-test behavior DD: AC vs ABC: p = 0.10). Taken together, the analysis of the punishment pattern revealed the expected parochial impact of group membership on altruistic norm enforcement, which is particularly pronounced in the DC condition when outgroup members defect against cooperating ingroup members.

fMRI data: Outgroup effects

In order to identify a functionally connected brain network which explains the strong impact of group membership on punishment behavior, we calculated in a first analysis the contrast *outgroup (BC) minus ingroup (AC+ABC) constellations* during the *DC condition*. The

resulting statistical parametric map (SPM, at $p < 0.005$, voxel extent threshold: 10 voxels, Lieberman and Cunningham 2009) mainly revealed increased activation in a hypothesized network of brain regions which have been shown to play key and functionally separated roles in punishment-related decision processes (de Quervain, et al. 2004; Knoch, et al. 2007; Knoch, et al. 2006; Kringelbach 2005; Sanfey, et al. 2003; Spitzer, et al. 2007), including the right orbitofrontal gyrus (rOFG, BA 11/47, x = 33, y = 39, z = -9), the right lateral prefrontal cortex (rLPFC, BA 44/45, x = 57, y = 12, z = 15), and the right dorsal caudatus (x = 15, y = 24, z = 9; Fig. 3, Supp. Table S1). Importantly, all these a priori regions of interests also survive small volume (SV) family-wise-error (FWE) corrections at $p < 0.05$ (see method section for details), except the dorsal caudatus which just falls short of the threshold with $p = 0.057$.

In order to corroborate the specificity of the findings in the rOFG, rLPFC, and right caudatus with respect to both lateralization pattern and condition effects, we conducted further analyses using either functional or spherical ROIs (see method section for details). First, we took a closer look at the lateralization pattern. We found no effect of group membership (BC vs AC, BC vs ABC, AC vs ABC) in the left hemisphere (paired t-tests in the DC condition: rOFG: all $p > 0.23$; rLPFC: all $p > 0.88$; right caudatus: all $p > 0.55$). Furthermore, these three brain regions were significantly more engaged in the right than in the left hemisphere in the DC condition (paired t-tests: OFG: $p = 0.001$; LPFC: $p = 0.05$; caudatus: $p = 0.009$), suggesting that the punishment-related activation in the OFG, LPFC, and caudatus are confined to the right hemisphere.

If the rOFG, rLPFC, and right caudatus are indeed involved in the decision-making process which leads to an increased punishment of defective outgroup member, then these brain regions should not show a differential effect of group membership for player A's cooperative decisions (behavioral patterns CC, CD), when third-parties' punishment behavior is virtually absent. Consistent with these assumption, the interaction effect of group constellation × behavior was significant in all these regions (all $p < 0.01$), and we found no effect of group membership (BC vs AC, BC vs ABC, AC vs ABC) on brain activity during these cooperative decision patterns (paired t-tests: rOFG: all $p > 0.46$; rLPFC: all $p > 0.11$, right caudatus: all $p > 0.46$). Furthermore, and in line with the markedly reduced differences in punishment behavior in the DD condition (see Fig. 2), all but one differential effect of group membership disappeared during the DD condition in these regions (all $p > 0.23$). The only exception

concerns the rLPFC which still shows enhanced activity ($p = 0.002$) in the outgroup constellation (BC) compared to one of the ingroup constellations (AC).

So far, we have shown that the rOFG, rLPFC, and right caudatus show a very similar and highly specific activity pattern, which is restricted to the right hemisphere and only present when third-parties strongly punish outgroup members who defect against cooperating ingroup members. Such an activity pattern suggests, but does not yet provide evidence, that these regions actually form a functionally connected network orchestrating punishment behavior. In order to provide this evidence, we applied Physio-Physiological Interaction (PPI) analyses (Friston, et al. 1997), which elucidate the influence that one neuronal system exerts over another (termed effective connectivity, see method section for details). Findings revealed (at $p < 0.005$, voxel extent threshold: 10 voxels) that activity in the rOFG positively modulates the effective connectivity between the right caudatus and rLPFC. In other words, we found that the higher the activity in the rOFG, the stronger the connectivity between right caudatus and rLPFC (Fig. 4A, for additional information and illustration please see Supp. Discussion S1 and Supp. Fig. S3A) – a finding that provides evidence that these three brain regions actually form a functionally connected neural network orchestrating punishment behavior. This evidence was further corroborated by the observation (at $p < 0.005$, voxel extent threshold: 10 voxels) that a neighboring part of the rLPFC (BA 45/46, $x = 48$, $y = 21$, $z = 21$) shows a positive connectivity pattern with the rOFG which depends on the punishment level. This means that the stronger third-parties punish defecting outgroup members in the DC condition, the stronger is the positive connectivity between the rOFG and the rLPFC (Fig. 4B).

fMRI data: Ingroup effects

Next, we reversed the contrast of the first analysis and wondered whether we find increased activity in the *ingroup (AC+ABC) minus the outgroup (BC) constellations* during the *DC condition*. Such a differential finding in the brain would be a first step in understanding the neural processes behind the phenomenon in which ingroup members incur much less severe punishment than outgroup members for the same defective behavior. We primarily found increased activity (at $p < 0.005$, voxel extent threshold: 10 voxels, Lieberman and Cunningham 2009) in three brain regions that are well-known to form a neural network involved in mentalizing processes (Gallagher and Frith 2003; Rilling, et al. 2004; Van Overwalle 2009), including dorsomedial prefrontal cortex (DMPFC, BA 9, $x = 6$, $y = 54$, $z = 30$) and bilateral temporo-parietal junction (TPJ, BA 39/40/22, left TPJ: $x = -45$, $y = -60$, $z =$

21; right TPJ: x = 57, y = -60, z = 30; Fig. 5, Supp. Table S1). All these a priori regions of interests survive small volume family-wise-error (FWE) corrections at $p < 0.05$ (see method section for details).

In order to corroborate the specificity of the findings in the DMPFC and bilateral TPJ, we conducted further analysis using functional ROIs in the DMPFC and spherical ROIs (5mm radius) in the bilateral TPJ (see method section for details). If this mentalizing network is involved in reducing the punishment of defecting ingroup members, then we should not observe a differential group effect in these brain regions when third-parties face cooperative decisions by player A (CC, CD). Consistent with this hypothesis, the interaction effect of group constellation × behavior was significant in all these regions (all $p < 0.01$) and we found no evidence for an impact of group membership (BC vs AC, BC vs ABC, AC vs ABC) on the activity pattern in these brain regions during cooperative decisions (paired t-tests: DMPFC: all $p > 0.24$, left TPJ: all $p > 0.32$, right TPJ: all $p > 0.41$).

Furthermore, we found that the functional connectivity (see method section) between two of the brain regions in this mentalizing network (DMPFC and left TPJ) depends on the third-parties' punishment levels of defecting ingroup members in the DC condition; the lower the third-parties' punishment level, the higher the functional connectivity between these two brain regions (at $p < 0.005$, voxel extent: 12 voxels, Fig. 6). In other words, the stronger the interaction between these two regions of the mentalizing network, the less ingroup members are punished for defecting against cooperating outgroup members. This finding provides additional evidence that the mentalizing network is indeed recruited in order to reduce the punishment of defecting ingroup members.

fMRI data: Functional connectivity analysis between the mentalizing network and the punishment network

So far, we have identified a functionally connected punishment-network in our analyses consisting of the rOFG, rLPFC, and right caudatus. This network shows enhanced activity and connectivity whenever third-parties strongly punish outgroup members who defect against cooperating ingroup members. On the other hand, third-parties punish ingroup members far less for the same defective behavior, which is associated with increased activity and connectivity in main areas of the mentalizing network consisting of the DMPFC and bilateral TPJ. In a next analysis, we examined whether we find evidence in our data that these two

neuronal networks are functionally connected. We hypothesize that the mentalizing network might modulate the activity in parts of the punishment network, thus enabling third-parties to implement a reduced punishment behavior for defecting ingroup members. We focused our analysis in particular on the orbitofrontal gyrus because we argue that third-parties observing defecting ingroup members might start a justification process in the mentalizing network, which affects the evaluation process assumed to take place in the OFG (Kringelbach 2005; Liu, et al. 2007; Rangel, et al. 2008). In order to answer this question, we applied Physio-Physiological Interaction (PPI) analyses (Friston, et al. 1997) using two areas of the mentalizing network as seed regions (DMPFC and left TPJ, see method section for details). Findings revealed (at $p < 0.005$, voxel extent threshold: 10 voxels) in good agreements with our hypotheses that the activity in the left TPJ modulates the effective connectivity between the DMPFC and areas of the OFG and rLPFC. In particular, we found that the activity in the left TPJ *positively* modulates the functional connectivity between the DMPFC and medial areas of the OFG (mOFG; BA 10/11, x = -15, y = 45, z = -9) known to be involved in positive evaluation processes (Kringelbach 2005; Liu, et al. 2007; Rangel, et al. 2008). In stark contrast, activity in the left TPJ *negatively* modulates the functional connectivity between the DMPFC and lateral areas of the OFG (BA 10/11, left: x = -42, y = 54, z = -9, right: x = 24, y = 60, z = -6) and rLPFC (BA 45/46, x = 54, y = 36, z = 15) (Fig. 7A/B; for additional information and illustration please see Supp. Discussion S1 and Supp. Fig. S3B-D). In other words, the higher the activity in the left TPJ, the stronger is the positive connectivity between DMPFC and mOFG, and the stronger is the negative connectivity between the DMPFC and lateral areas of the OFG and rLPFC. Notably, these negative connectivity effects are localized in neighboring and overlapping areas of the OFG and rLPFC shown to be involved in the punishment-related decision process in the DC condition (depicted in Fig. 3 and 4). Taken together, these findings support our hypothesis that the mentalizing network might control the activity in the punishment-network, in particular by affecting the evaluation of the ingroup member's defective behavior. As a consequence, this reduced negative and/or increased positive evaluation might enable third-parties to implement a reduced punishment for the same defective behavior.

Discussion

This study is the first to explore the neural networks involved in orchestrating the parochial nature of altruistic norm enforcement – a pervasive psychological phenomenon, which has shaped the human society in decisive ways (Bernhard, et al. 2006; Choi and Bowles 2007; Efferson, et al. 2008). Findings revealed that third-parties' parochial punishment pattern is modulated by functionally connected neural networks previously shown to be involved in negative and positive evaluation processes (VLPFC, VMPFC, OFG), the weighting of economic self-interests (rLPFC), the appetitive and motivational component of punishment (right dorsal caudatus), and mentalizing processes (DMPFC, bilateral TPJ, for a summary of the neural findings please see Fig. 8). In the following, we will discuss the neural findings in detail.

The increased punishment of outgroup members who defect against cooperating ingroup members was associated with increased activity in two hypothesized areas in the right lateral PFC and right dorsal caudatus, which have been shown to play important and functionally distinctive roles during the implementation of second-party punishment decisions where the norm violation directly affects the second-party punisher (in contrast to the third-party punisher). Existing studies (Knoch, et al. 2007; Knoch, et al. 2006; van 't Wout, et al. 2005) have demonstrated that disrupting the function of the control-related area of the right lateral PFC reduces subjects' willingness to punish a norm violation committed intentionally, showing that the area is causally involved in costly punishment behavior. Crucially, the fairness judgment remained unaffected by the disruption of the PFC, suggesting that the subjects' ability to identify and emotionally experience the norm violations were not compromised. These findings thus imply that the subjects are less able to implement costly punishment following disruption of the function of the right lateral PFC, due to the lack of prefrontal-mediated modulation of economic self-interests. On the other hand, the right dorsal striatum, a brain area strongly implicated in reward processing (Delgado, et al. 2003; Fliessbach, et al. 2007; Liu, et al. 2007; O'Doherty, et al. 2004), has also been demonstrated to be activated during second-party punishment decisions (de Quervain, et al. 2004), consistent with the view that this brain area motivates and reinforces the punishment act. Taken together, the increased activity of these two brain regions during both second and third-party norm enforcement processes indicates, as hypothesized, that the implementation of both second and third-party punishment decisions rely on similar neural circuitry. This interpretation is consistent with the findings of a study on third-party punishment judgments of fictive norm

violations (Buckholtz, et al. 2008), where increased activity in the right lateral PFC was found when participants made judgments about the appropriate punishment for norm transgressions committed intentionally (rather than unintentionally).

We hypothesized that parochial punishment patterns require the activation of further processes in the third-parties' brains because it is otherwise difficult to explain why third parties punish the very *same* norm violation much more severely when an outgroup member (as opposed to an ingroup member) commits the transgression. In particular, we conjectured that two key processes must take place in third-party brains: First, a differential evaluation of outgroup and ingroup members' defective behavior and, second, a mentalizing process which justifies this differential evaluation process. We predicted that these additional mental processes will yield differential activity and connectivity patterns across specific brain regions that have been shown to be involved in (1) evaluative processes and (2) in mentalizing processes in previous studies.

The first prediction is supported by the observation of increased activity in the rOFG whenever third-parties severely punished outgroup members who defect against cooperating ingroup members and strongly reduced activity in the same brain region whenever ingroup members were punished far less for the very same norm violation. The lateral OFG, and in particular the rOFG, are brain areas shown to be associated with negative evaluation processes (Kringelbach 2005; Liu, et al. 2007; Rilling, et al. 2007; Seymour, et al. 2007; Spitzer, et al. 2007). The strongly increased activation of rOFG during the punishment of outgroup defectors is consistent with the view that third-parties evaluate outgroup members' defective behavior more negatively than ingroup members' defection. Moreover, we also find that the evaluation-related area in the rOFG mediates the effective connectivity between the two other areas in the rLPFC and rCaudatus that are activated whenever third-parties strongly punish outgroup defectors. These findings suggest that the stronger the negative evaluation of outgroup defection, and the stronger the associated activation of rOFG, the stronger the functional connectivity between the two areas known to be involved in implementing the punishment behavior. In other words, a strong negative evaluation of outgroup members' defective behavior might trigger a cascade of functionally connected neural processes which enable third-parties to implement a more costly punishment by down-weighting their own economic self-interest and motivating the punishment act. Further evidence for this interpretation is provided by the observation that the strength of punishment of defecting

outgroup members depends on the functional connectivity between the rOFG and rLPFC. The stronger these areas are functionally connected, the stronger outgroup members are punished for committing a norm violation against cooperating ingroup members.

The data also supports our second prediction that the much lower punishment of ingroup members is associated with activation of brain regions associated with mentalizing processes because we find increased activity in the DMPFC and bilateral TPJ when the third-parties face an ingroup defector (compared to an outgroup defector). If third-parties attempt to understand the intentions or goals behind ingroup members' defective behavior because they try to find mitigating reasons that provide an excuse for ingroup defectors, we would expect the activation of DMPFC and TPJ because these two brain regions are key components in the mentalizing network known to be involved in inferring temporary goals, intentions, desires, as well as more enduring dispositions of others (Gallagher and Frith 2003; Hampton, et al. 2008; Mitchell, et al. 2005; Steinbeis and Koelsch 2009; Van Overwalle 2009). Our connectivity findings both within the mentalizing network and between the mentalizing and the punishment network corroborate the hypothesis that the mentalizing regions modulate punishment behavior. *First,* the functional connectivity of two areas of the mentalizing network (DMPFC and left TPJ) depends on third-parties' punishment decisions. More precisely, the less third parties punish ingroup defectors, the stronger is the functional connectivity between these two key areas of the mentalizing network. *Second,* the same two areas of the mentalizing network (DMPFC and left TPJ) show a connectivity pattern with areas of the punishment network, suggesting that mentalizing-related activity reduces the punishment behavior by modulating the activity in evaluation-related (rOFG, mOFG) and control-related areas of the lateral PFC (rLPFC). More precisely, the neural data suggest that the left TPJ *negatively* modulates the effective connectivity between the DMPFC and lateral areas of the PFC (rOFG, rLPFC), whereas the left TPJ *positively* modulates the effective connectivity between the DMPFC and medial areas of the OFG known to be involved in positive evaluation processes (Hare, et al. 2009; Kringelbach 2005). Thus, the lower punishment of ingroup members' defective behavior may have been implemented via the modulatory role of left TPJ on evaluation related prefrontal areas such as mOFG and rOFG.

The lack of mentalizing-related brain activity and the associated increased punishment in situations when outgroup members commit the norm transgression are particularly interesting in light of a recent publication (Harris and Fiske 2006). In this study, participants saw images

of different outgroups which varied in dimensions of competence (high or low) and warmth (high = friend, low = foe). Crucially, the mentalizing network only failed to be activated if participants faced extreme outgroups, i.e. people who were both low in competence and low in warmth (e.g. homeless people, drug addicts). The authors interpreted this lack of mentalizing-related activity as an indication that extreme outgroups may not be perceived as fully human, may even be dehumanized by denying characteristics to them that are uniquely human (representing them as animal-like) and those that constitute human nature (representing them as objects or automata) (Allport 1954; Haslam 2006). This lack of activity in areas of the mentalizing network in both studies permits the speculation that the defecting outgroup members of our paradigm are treated similarly to the extreme outgroup members in the study by Harris and colleagues. In this context, it is important to remember that our outgroup does not consists of extreme outgroup members such as homeless people or drug addicts. Instead, the characteristics of the outgroup and ingroup members are identical; all are officer candidates in the Swiss army, there are no significant differences in education or age, and the assignment to the different platoons/groups was random. Thus, the finding that the extreme outgroups (e.g. homeless people and drug addicts) and the defecting outgroup members of our study evoke a similar activity pattern in the mentalizing network (i.e. no increased activity compared to baseline) further illustrates the strong impact of parochial altruism on neural activations.

Finally, we would like to point out that some of our interpretations (e.g. negative evaluations in the rOFG and mentalizing in the DMPFC and bilateral TPJ) rely on assumptions about specific cognitive functions subserved by these brain regions in the type of behavioral paradigm we implemented. If we had complemented our neuroimaging and behavioral punishment data with subjective ratings or judgments, these assumptions could have been further strengthened. For example, if we had asked subjects about their mentalizing processes, we would have been able to clarify more precisely whether the increased activity in the mentalizing network indicates an overabundance of mentalizing leading to increased justification for in-group members or rather a deficit in mentalizing for outgroup members leading to decreased justification. However, despite this limitation of the study, we would like to emphasize that no assumptions about specific cognitive functions subserved by these brain regions are necessary to render the results of our paper interesting and important because no other paper has yet identified the activity and connectivity patterns associated with parochial norm enforcement (please check summary of the results below).

Summing up, this study is the first to reveal the neural circuitry of the impact of parochial altruism on social norm enforcement. In order to study this prevalent psychological phenomenon, we applied a third-party punishment paradigm with manipulation of group membership using real social groups, the involvement of real monetary stakes, and the requirement to curb immediate self-interests in order to enforce social norms. We found **(1)** that third-parties' higher punishment of defecting outgroup members is associated with increased activity and connectivity in a functionally connected neural network involved in punishment-related decision processes, including the rOFG, rLPFC, and right dorsal caudatus. Furthermore, **(2)** the functional connectivity between two areas of this punishment network predicts third-parties' punishment of defecting outgroup members. More precisely, the stronger the rOFG and rLPFC are functionally connected, the more severely outgroup members are punished for defecting cooperating ingroup members. In sharp contrast, **(3)** the lower punishment of defecting ingroup members is associated with increased activity in brain areas well-known to be involved in mentalizing processes (including the DMPFC and bilateral TPJ), as if third-parties tried to understand the underlying intentions and goals behind ingroup members' defective behavior. The conjecture that the mentalizing network modulates punishment is corroborated by the finding that **(4)** the functional connectivity between two areas of this mentalizing network (DMPFC and left TPJ) is related to the punishment of ingroup defectors in a particular way: the stronger the connectivity between these two areas, the less third-parties punish ingroup defectors. Finally, **(5)** the functional connectivity between areas of the mentalizing (DMPFC and left TPJ) and punishment network suggests that the reduction in punishment is associated with the modulation of the neural activity in the right orbitofrontal gyrus and right lateral prefrontal cortex, i.e. in the same areas showing enhanced activity and connectivity whenever third-parties strongly punished defecting outgroup members. The neural findings of this study thus provide evidence for the view that humans have developed elaborate neural circuitry for social cognition that modulates the parochial nature of altruistic norm enforcement – a prevalent psychological phenomenon that has shaped humans' cooperative, altruistic, and punishment-related behavior in decisive ways.

Methods

Subjects

A total of 16 healthy, right-handed male subjects (mean age ± s.d., 24.5 ± 2.2, max: 27, min: 20) participated in the fMRI-study. None of the participants had to be excluded from the analyses. All subjects were free of chronic diseases, mental disorders, medication, and drug or alcohol abuse. The study was carried out in accordance with the Declaration of Helsinki principles and approved by the institutional ethics committee. All subjects gave written, informed consent and were informed of their right to discontinue participation at any time.

Group manipulation: Real social groups

All 16 subjects who took part in the fMRI experiment were completing a four-week phase of officer training in the Swiss army at the time of the experiment. This training course brings officer candidates from all branches of service together to the same location in order to promote exchanges of perspective among different branches of service. Training involves mainly coursework on principles of security, combat in large military units, logistics, and leadership. Important for the investigation of the impact of group membership on social norm enforcement, the officer candidates are randomly assigned to platoons (groups) at the beginning of the training course. This random assignment mechanism is ideal for the experimental paradigm in several ways (for details see Goette, et al. 2006). *First,* trainees know that platoon composition is identical and that none of the officers could choose which platoon to join. Statistical tests reveal no significant differences in platoon composition with respect to branch of service, education, or age. *Second,* there is no competition between the groups for evaluations or other resources. *Third,* despite random assignment to platoons, social ties form very quickly. Officers indicated in a questionnaire that they spent significantly more time off duty with members of their own platoon. Thus, the officer candidates interact almost exclusively with members of their own platoon, both during on-duty and off-duty time.

Design

We applied a third-party punishment paradigm (Fehr and Fischbacher 2004b) with group manipulation (Bernhard, et al. 2006; Goette, et al. 2006) consisting of two decision stages, one conducted in the training course of the Swiss army and one conducted in the fMRI scanner. The subjects knew in all stages of the experiment that there were no repeated interactions and that all interactions were conducted in complete anonymity. During the first

decision stage conducted in the training course of the Swiss army, the officer candidates played (in the role of player A and B) a simultaneous Prisoners' Dilemma game, where they could decide either to cooperate or defect. Thus, 4 behavioral patterns are possible: player A and B cooperate (CC), player A and B defect (DD), player A cooperate and player B defects (CD), and player A defects and player B cooperates (DC). Player A and B knew that other officer candidates from the same training course would be confronted with their decisions, and that they would receive the opportunity to judge their behavior in the role of a third-party (player C) by assigning deduction/punishment points. In total, 16 officer candidates were invited to take part in this second decision stage conducted in the fMRI-scanner, where they were confronted with 30 decisions stemming from different players A and B. At the beginning of each punishment trial, they received an endowment of 10 points which they could either keep or use to punish player A. 1 point assigned for punishment reduced the punished player's income by 3 points. Points not used for punishment were exchanged into real money and paid to player C at the end of the experiment (for details on exchange rate please see below). Player C was always informed about the group affiliation of player A and B, that is whether the players were from his own or another platoon (in-/outgroup manipulation). There are 3 group constellations in the experiment; two ingroup constellations (ABC, AC), where player A (who can be punished by C) is from his own platoon and one outgroup constellation (BC), where player A (who can be punished by C) is from another platoon.

Subjects were paid according to their decisions (player A, B, and C) and the decisions of their interaction partners (player A and B). While players A and B received their money per mail shortly after the scanning session was conducted, player C was immediately paid at the end of the scanning session. The exchange rate was as follows: 10 points = 2 Swiss Francs, that is about $ 2.

Subjects read written instructions describing the details of the paradigm, including the payoff rules, prior to both decision stages in the Swiss army training course and in the fMRI scanner. After the subjects had read the instructions, we checked whether they understood the payoff rules and the treatment conditions by means of several hypothetical questions. All subjects answered the control questions correctly.

Procedure in the scanner

The computer screens that the third-parties needed to see during the punishment trials (see Fig.1 for an example) were presented with a video projector onto a translucent screen that subjects viewed inside the scanner via a mirror. On these judgment screens, players A and B were represented by schematic pictures of two humans whose coloring in grey or black indicated the players' group affiliation. In half of the subjects, grey indicated an ingroup member and black an outgroup member, while the coloring for ingroup and outgroup members was reversed for the other half of the subjects. In addition to the coloring, the players' group affiliation was depicted verbally (your group/other group). Finally, the players' behavior, that is whether these players defected (kept the points) or cooperated (transferred the points) was also depicted verbally. After this information on group affiliation and behavior was presented for 4 seconds, 4 buttons representing the four punishment options were presented on the same screen, indicating that the subjects could now implement their punishment decisions, by means of a 4-button input device. In half of the subjects, these punishment options were presented with a scale ranging from 0 to 9 deduction points (0/3/6/9) and in the other half of subjects, the punishment scale was reversed (9/6/3/0). On average, punishment decisions were implemented 6.92 seconds (standard error: 0.36) after the onset of the judgment screen. After pressing the button, the judgment decision remained on the screen for another two seconds and was then replaced by a fixation cross, which separated the punishment trials by 14 seconds.

The software package z-Tree (Fischbacher 2007), a program for conducting behavioral experiments in combination with neuroimaging, was used for presenting screens and for collecting behavioral and timing data.

Behavioral analyses

In order to analyze third-parties' punishment levels, we first calculated an average punishment level for each condition (3 group constellations × 4 behavioral patterns). We then used the statistical software package SPSS 15 for PC (SPSS Inc., Chicago, IL) for the different analyses of the behavioral data. Please see results section for details about the statistical tests conducted, including paired t-tests, simple t-tests, and repeated measures ANOVA with within-subjects factor group constellation (BC, AC, ABC) and behavioral patterns (CC, CD, DC, DD). Results were considered significant at the level of $p < 0.05$ (two-tailed). In case of a

significant multivariate effects, post hoc paired t-tests were computed using the Bonferroni correction according to Holm (Holm 1979). As effect size measure ETA^2 is reported.

fMRI analyses: Image acquisition

The experiment was conducted on a 3 Tesla Philips Intera whole body MR Scanner (Philips Medical Systems, Best, The Netherlands) equipped with an 8-channel Philips SENSE head coil. Structural image acquisition consisted of 180 T1–weighted transversal images (0.75 mm slice thickness). For functional imaging, a total of 280 volumes were obtained using a SENSitivity Encoded (SENSE; (Pruessmann, et al. 1999)) T2*-weighted echo-planar imaging sequence with an acceleration factor of 2.0. 40 axial slices were acquired covering the whole brain with a slice thickness of 3mm; no inter-slice gap; interleaved acquisition; TR = 3000 ms; TE = 35ms; flip angle = 77°, field of view = 220mm; matrix size = 128 × 128. We used a tilted acquisition in an oblique orientation at 30° to the AC-PC line in order to optimize functional sensitivity in orbitofrontal cortex and medial temporal lobes.

fMRI analyses: Preprocessing

For the preprocessing and statistical analyses, the statistical parametric mapping software package (SPM5, Wellcome Department of Cognitive Neurology, London, UK) implemented in Matlab (Version 7) were used. For analysis, all images were realigned to the first volume, corrected for motion artifacts and time of acquisition within a TR, normalized (3 × 3 × 3 mm^3) into standard stereotaxic space (template provided by the Montreal Neurological Institute), and smoothed using an 8 mm full-width-at-half-maximum Gaussian kernel. A band-pass filter composed of a discrete cosine-basis function with a cut-off period of 128 seconds for the high-pass filter was applied. In order to increase signal to noise ratio, global intensity changes were minimized by scaling each image to the grand mean.

fMRI analyses: General linear model (GLM)

We performed random-effects analyses on the functional data for the punishment period. For that purpose, we defined a general linear model (GLM) with the following regressors of interests: 3 group constellations (BC, AC, ABC) × 4 behavioral patterns (CC, CD, DC, DD). The length of each of these regressors was individually modeled from the onset of the punishment trials until the subject's button press. All regressors were convolved with a canonical hemodynamic response function (HRF). The 6 scan-to-scan motion parameters

produced during realignment were included as additional regressors in the SPM analysis to account for residual effects of scan to scan motion.

For second-level random effects analysis, the single-subject Beta-estimates were entered into a repeated-measures of ANOVA with within-subject factor group constellations (BC, AC, ABC) and within-subject factor behavioral patterns (CC, CD, DC, DD). Due to the fact that the study was specifically designed to reveal the parochial nature of altruistic norm enforcement, we focused on the DC trials in our analyses of the brain activity pattern, where we observed, as expected, the strongest parochial punishment pattern. All other behavioral trials (CC, CD, DD) were primarily used in the analyses to demonstrate the specificity of the parochial activity pattern during the DC condition.

In order to reveal the neural underpinnings of the parochial punishment pattern, we focused on the following two brain contrast:

- *DC trials only: Outgroup (BC) minus Ingroup (AC+ABC)* ^{weighted}
- *DC trials only: Ingroup (AC + ABC)* ^{weighted} *minus Outgroup (BC)*

The first brain contrast allowed us to test our hypothesis that the increased punishment of defecting outgroup members is associated with increased activity in brain areas known to play key and functionally distinct roles in punishment-related decision processes (lateral OFG, lateral PFC, and caudatus). The second brain contrast allowed us to test our hypothesis that the reduced punishment of defecting ingroup members is associated with increased activity in key areas of the mentalizing network (DMPFC, TPJ):

In order to increase the specificity of the findings during these *DC trials only* contrasts, we exclusively masked them at $p < 0.05$ with the same group constellation contrast (Outgroup minus Ingroup and vice versa, respectively), but calculated with trials where player A (whom C can punish) shows a cooperative behavioral pattern (CC, CD). This masking procedure excludes all regions (at $p < 0.05$) that show an unspecific main effect of group membership. In other words, we report in these *DC trials only* contrasts brain areas, which are differentially activated solely when third-parties demonstrate a highly distinctive parochial punishment pattern.

Although not the main aim of the current manuscript, we calculated the following two additional contrasts in order to reveal the brain areas demonstrating a main effect of group

membership, including those trials where no (CC, CD) or only a marginal (DD) parochial punishment pattern was expected and found (see supporting analysis S1):

- *All behavioral trials: Outgroup (BC) minus Ingroup (AC+ABC)*^{weighted}
- *All behavioral trials: Ingroup (AC + ABC)* ^{weighted} *minus Outgroup (BC)*

fMRI-Analyses: Statistical inferences

We report results in a priori regions of interests [previously defined in neuroimaging studies on punishment (Buckholtz, et al. 2008; de Quervain, et al. 2004; Knoch, et al. 2006; Sanfey, et al. 2003; Spitzer, et al. 2007; Strobel, et al. 2011) and mentalizing (Van Overwalle 2009)]: OFG, right lateral PFC, caudatus, DMPFC, TPJ where activations are significant at $p < 0.005$ uncorrected for multiple comparisons with an extent threshold of 10 voxels (Lieberman and Cunningham 2009), and survive small volume corrections (SVC) for multiple comparisons (or family-wise error [FWE] corrections across the whole brain). The SVC procedure, as implemented in SPM5 using the FWE correction procedure ($p < 0.05$), allows results to be corrected for multiple non-independent comparisons with a defined region of interest. For the SVC procedure, we used anatomical masks (lateral OFG, caudatus) obtained from the WFU PickAtlas toolbox (Maldjian, et al. 2003), and 20 mm spheres centered on coordinates derived from previous work. For the areas of the mentalizing network, we used a recently published meta-analysis (Van Overwalle 2009) on social cognition to define the peaks in the lTPJ (x = -49, y = -58, z = 22) , rTPJ (x = 53, y = -54, z = 22) and DMPFC (x = -3, y = 48, z = 30), which consisted of the average coordinates of those mentalizing tasks (including goal, intention and trait inferences, and morality judgments) consistently activating these brain areas. For the punishment-related activated in the rLPFC, we averaged the peak coordinates of a second-party norm enforcement study (x = 40, y = 36, z = 22) (Sanfey, et al. 2003) where disruption by rTMS (Knoch, et al. 2006) or tDCS (Knoch, et al. 2007) reduces subject's ability to control their economic self-interest. Activations in other regions were only considered significant if they survived whole-brain FWE correction for multiple comparisons at $p < 0.05$ [in line with established procedures (Frackowiak, et al. 2004)], but are reported for completeness at a threshold of $p < 0.005$ uncorrected for multiple comparisons. Reported voxels conform to Montreal Neurological Institute (MNI) coordinate space. The right side of the brain is displayed on the right side in our illustrations.

fMRI-analyses: ROI analyses

In order to illustrate the specificity of the findings both with regard to group constellations (BC, AC, ABC), behavioral patterns (CC, CD, DC, DD), and lateralization effects, we created either functional or spherical ROIs using the MarsBaR software. Functional ROIs encompassed all voxels that were significantly ($p < 0.005$) activated in the corresponding contrast analyses, whereas spherical ROIs consisted of a 5 mm sphere around the peak of activity. The rule for applying spherical or functional ROI's was as follows: For regions with a voxel extent of 20 or more voxels (at $p < 0.005$), we created spherical ROIs (bilateral temporo-parietal-junction), whereas we created functional ROIs for all other regions with a voxel extent between 10 and 20 voxels (rOFG, rLPFC, rCaudatus, DMPFC). Please note that the conducted statistical analyses (based on extracted Beta-estimates) do not significantly change if we apply the same ROI type for all regions, either functional or spherical.

In order to corroborate the ROI analyses described above, additional region of interests analyses were performed on anatomical or spherical ROIs defined by prior studies. The advantage of this approach is that the definition of ROIs is independent of the findings in the present study and thus less biased. To conduct these ROI analyses, we applied the same independent ROIs as we had used for the small volume family-wise-error corrections (described in detail above, please see statistical inference). Importantly, the findings of the two ROI-analyses do not differ. In particular, the specificity of the parochial activity pattern reported for the DC condition was corroborated, since there was no impact of group membership on these regions during cooperative behavioral decisions (all $p > 0.25$).

fMRI-analyses: Physio-Physiological Interaction analyses

In order to reveal the functional connectivity between brain areas orchestrating the parochial nature of altruistic norm enforcement, we applied Physio-Physiological Interaction (PPI) analyses (Friston, et al. 1997), which elucidate the influence that one neuronal system exerts over another (termed effective connectivity). For that purpose, we extracted mean-corrected and high-pass filtered time series of the rOFG, the rCaudatus, the DMPFC, and the left TPJ from a 5 mm spherical ROI around the peak of activation derived from the *DC trials only* contrasts. Once these time series were obtained for each subject, the interaction term (referred to as "PPI regressor") was computed as the vector resulting from the element-by-element product of two mean corrected time series. Based on our hypotheses and the findings of the GLM analyses, we created the following two interaction terms: interaction term of the time

series of rOFG and rCaudatus as well as of the time series of left TPJ and DMPFC. We used these two interaction terms as regressors of interests in two independent first level analyses, with the two single time series included as regressors of non-interests (either rOFG and rCaudatus or DMPFC and left TPJ). Each subject's Beta-estimates of the two PPI regressors were then taken to random-effects group analyses and entered into two one-sample t-tests. We were particularly interested in answering the following two questions in our analyses of these two PPI-regressors: First, do we find evidence in our data that the evaluation-related area of the rOFG *positively* modulates the connectivity between the rCaudatus and rLPFC, two areas thought to be critically involved in implementing punishment-related decision processes? Second, do we find evidence that one part of the mentalizing network in the left TPJ *negatively* modulates the connectivity between the other part of the mentalizing network in the DMPFC and the areas involved in punishment-related decision processes, including the rOFG, rLPFC, and the rCaudatus?

Finally, we examined whether differences in functional connectivity within areas of the mentalizing and punishment network, respectively, can explain the individual differences in punishment behavior during the DC condition. For that purpose, we entered the Beta-estimates from the single time-course regressors of the rOFG and DMPFC (included as regressor of non-interests in the PPI analyses described above) into two multiple regression analyses. These regression analyses included the individual punishment levels from the DC condition as covariates. More precisely, we used third-parties' individual punishment level of defecting outgroup members to search for areas within the punishment network whose connectivity with the rOFG depends on the individual punishment level. In contrast, we used third-parties' individual punishment levels of defecting ingroup members to search for areas within the mentalizing network, where the connectivity with the DMPFC depends on the individual punishment level. In the former case, we expected a positive correlation within the punishment network, whereas we expected a negative correlation within the mentalizing network in the latter case.

Due to strong a priori hypotheses, the significant thresholds for all connectivity analyses were set at $p < 0.005$ with a cluster extent threshold of 10 voxels (Lieberman and Cunningham 2009). For illustrative purposes, we created functional ROIs using the MarsBaR software by selecting all voxels that were significantly activated at $p < 0.005$ together with a cluster extent threshold of 10 voxels in the corresponding analyses.

ACKNOWLEDGMENTS

This work is part of Project 9 of the National Competence Center for Research (NCCR) in Affective Sciences. The NCCR is financed by the Swiss National Science Foundation. E.F. also gratefully acknowledges support from the research priority program at the University of Zurich on the "Foundations of Human Social Behavior".

Figures 1-8

Figure 1: Design and decision screen.

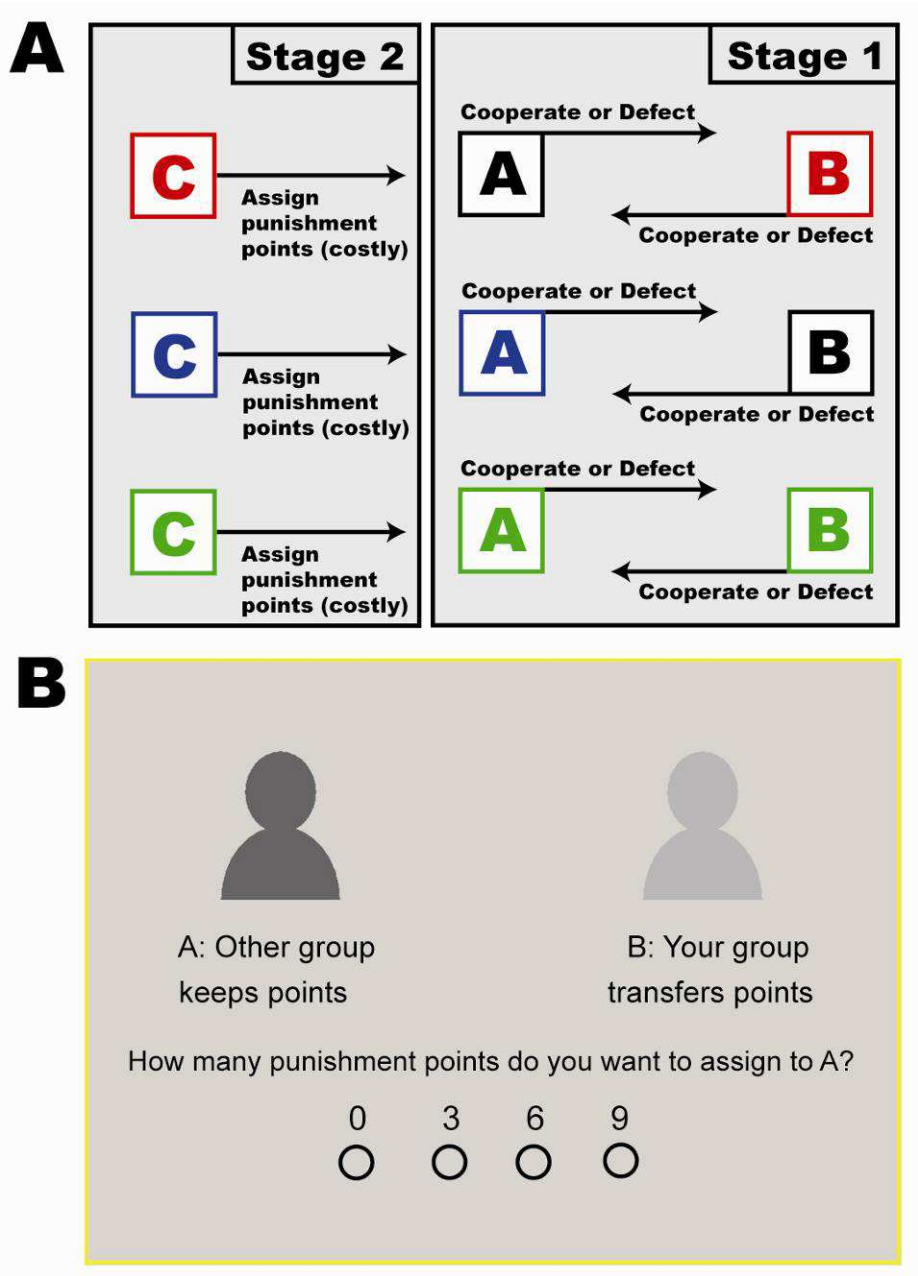

Figure 1: Design and decision screen. (A) Depicted is the third-party punishment paradigm with group manipulation (ingroup/outgroup). During stage 1, which took place in the training course of the Swiss army, player A and player B (officer candidates) played a simultaneous Prisoners' Dilemma Game (PD), in which they were free to decide whether to cooperate (transfer the points) or to defect (keep the points). During stage 2, which took place in the fMRI-scanner, some of the officer candidates in the role of a third-party (player C) were confronted with the decisions of player A and B and had the opportunity to assign (costly) punishment points to one of the players. For that purpose, player C was endowed with 10 points for each judgment trial. 1 point assigned for punishment reduced the income of the punished player by 3 points. Note that we recoded all of player C's decisions in such a way

that A always refers to the player that C *can* punish, whereas B always refers to the player that C *cannot* punish. Crucially, player A (whom C could punish) was either from the same group/platoon as player C, as in the group constellations ABC (in green color) and AC (in blue color), or was from a different group/platoon as in the group constellation BC (in red color). Thus, ABC and AC are ingroup constellations, whereas BC is an outgroup constellation. **(B)** Depicted is an example for a decision screen third-parties saw during the scanning session. In this particular case, third-parties were confronted with the outgroup constellation BC and an outgroup member who defected against a cooperating ingroup member. The group affiliation of player A and B was indicated both verbally (your group/other group) and schematically (in black or grey color). Please note that we reversed the color for the schematic depiction of ingroup and outgroup members and the punishment scale (9 6 3 0 instead of 0 3 6 9) for half of the subjects.

Figure 2: Punishment behavior.

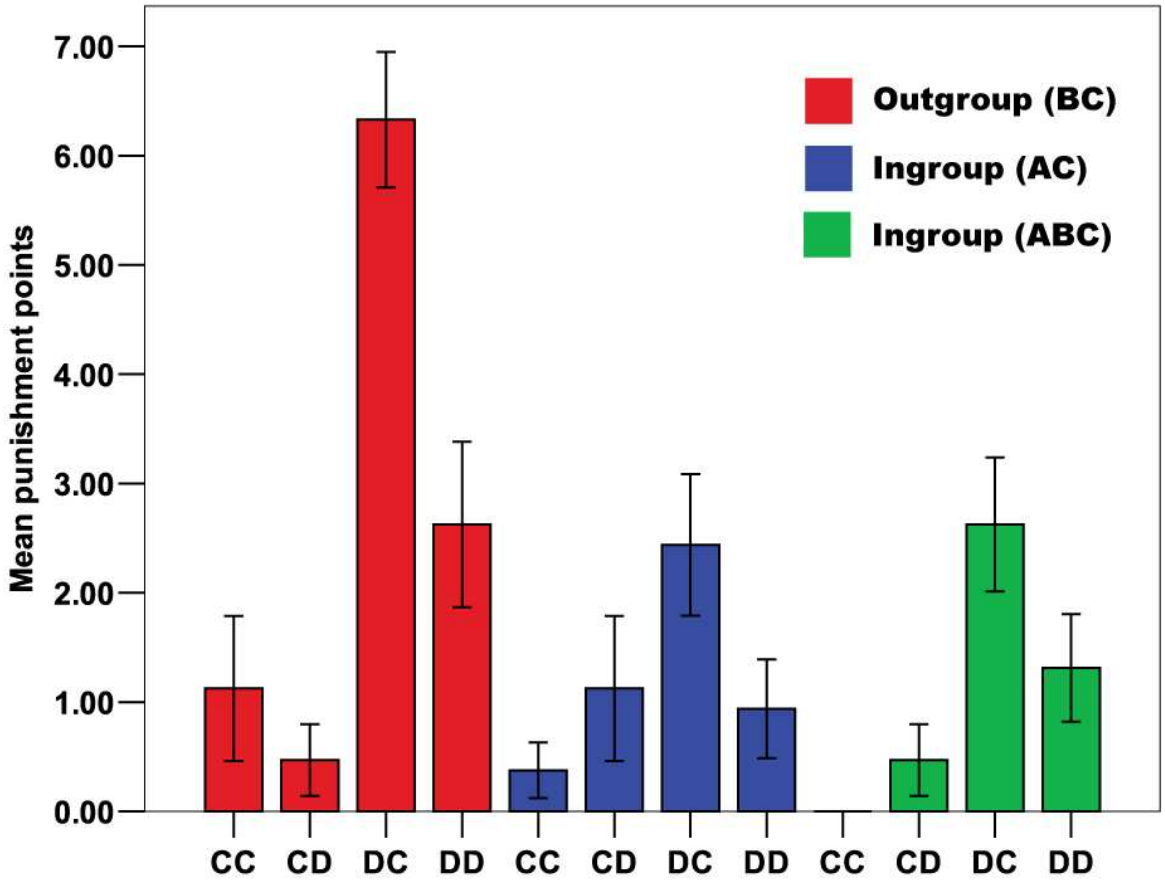

Figure 2: Punishment behavior. Analyses of third-parties' punishment behavior of player A revealed, as expected, a strong impact of group membership on punishment behavior when player A defected and player B cooperated (DC), a weak effect when both players defected (DD), and no significant impact of group membership when player A cooperated (CC, CD). Thus, third-parties' punishment behavior revealed the expected parochial pattern, qualified by increased punishment of outgroup members and reduced punishment of ingroup members for the same defective behavior.

Figure 3: Outgroup effects: Punishment network.

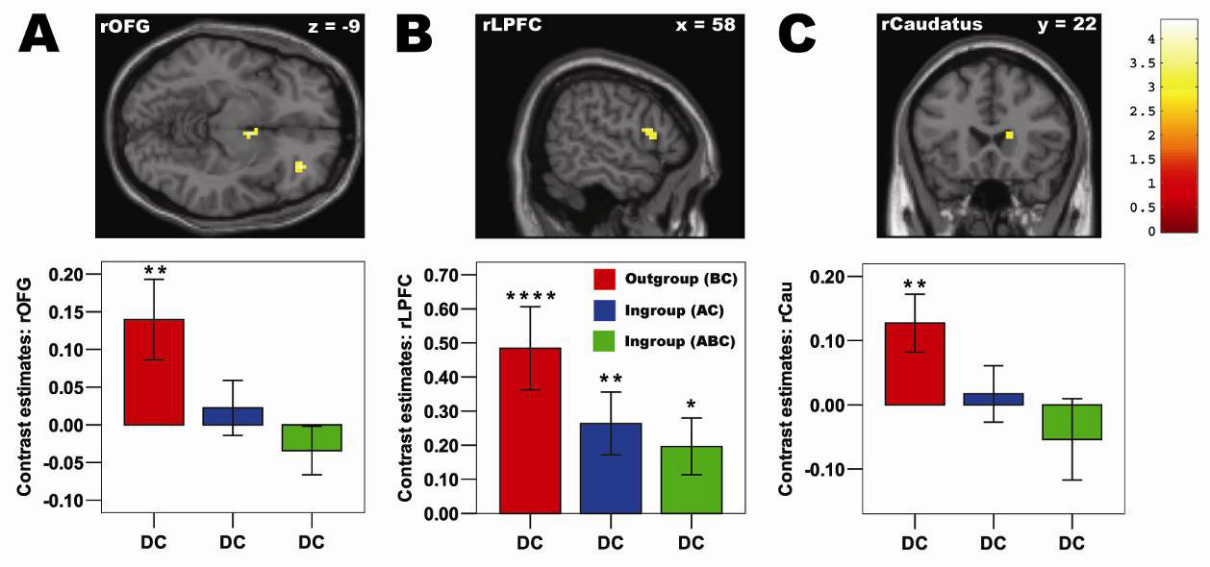

Figure 3: Outgroup effects: Punishment network. Depicted is the increased activation in the brain (at $p < 0.005$, voxel extent threshold: 10 voxels, activity in all regions survives small volume family-wise-error (FWE) corrections at $p < 0.05$, except the dorsal caudatus which just falls short of the threshold with $p = 0.057$, see methods section for details) contrasting the outgroup (BC) minus the ingroup (AC+ABC) constellations, when player A defected and player B cooperated (behavioral pattern DC). Consistent with the increased punishment pattern in the outgroup condition (BC), increased activity was mainly found in brain areas involved in punishment-related decision processes, including **(A)** right orbitofrontal gyrus (BA 11/47, x = 33, y = 39, z = -9), **(B)** right lateral prefrontal cortex (BA 44/45, x = 57, y = 12, z = 15) , and **(C)** right dorsal caudatus (x = 15, y = 24, z = 9). Bar plots representing contrast estimates (in/outgroup vs baseline) of functional ROIs (see method section for details) revealed in accordance with the similar punishment pattern that the two ingroup constellations (AC, ABC) show a highly similar activity pattern ($p > 0.25$ for all paired t-tests between the two ingroup constellations). Asterisks denote increased activity compared to baseline at $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.005$ (***) or $p < 0.001$ (****). Please see supporting Figure S1 for event-related BOLD time courses of the depicted brain regions.

Figure 4: Connectivity analyses within the punishment network.

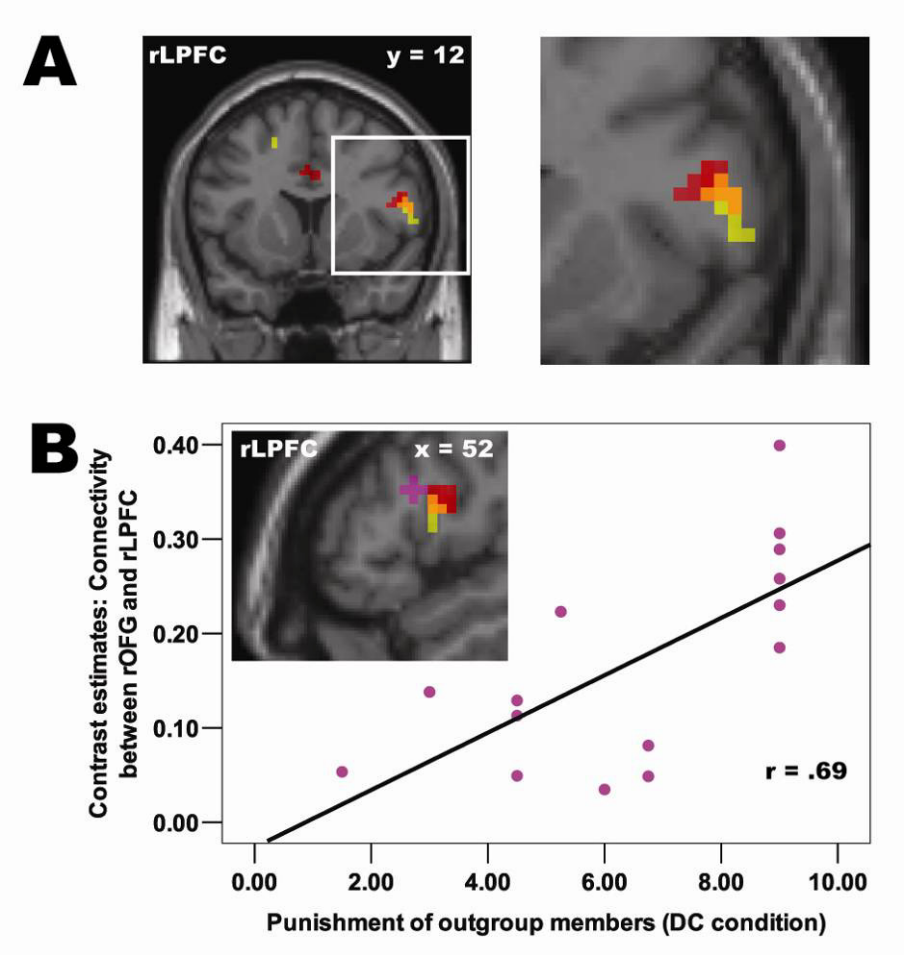

Figure 4: Connectivity analyses within the punishment network. (A) Physio-Physiological Interaction (PPI) analyses using the rOFG and right caudatus as seed regions revealed evidence (at $p < 0.005$, voxel extent threshold: 10 voxels) that the rOFG positively modulates the functional connectivity between right caudatus and rLPFC – notably in the same area of the rLPFC showing increased activity when third-parties strongly punish outgroup members who defect against cooperating ingroup members (the same activation as illustrated in Fig. 3B). These highly specific connectivity and activity patterns provide evidence for a functionally connected neural network orchestrating punishment behavior. Color coding: connectivity effect depicted in red, activation level effect depicted in yellow, overlap depicted in orange. **(B)** Connectivity analyses using the rOFG as seed region revealed (at $p < 0.005$, voxel extent threshold: 10 voxels, in violet color) that the functional connectivity between the rOFG and rLPFC depends on third-parties' punishment level, that is the higher third-parties punish defecting outgroup members in the DC condition, the stronger is the functional connectivity between these two regions. The scatter plot depicts this effect using a functional ROI of the rLPFC (BA 45/46, x = 48, y = 21, z = 21). In order to visualize the spatial proximity of all activation and connectivity effects in the rLPFC, the same activity and connectivity patterns described in (A) are also depicted here in (B) in the same colors. For display purposes, all activation and connectivity patterns in (A) and (B) are depicted at $p < 0.01$.

Figure 5: Ingroup effects: Mentalizing network.

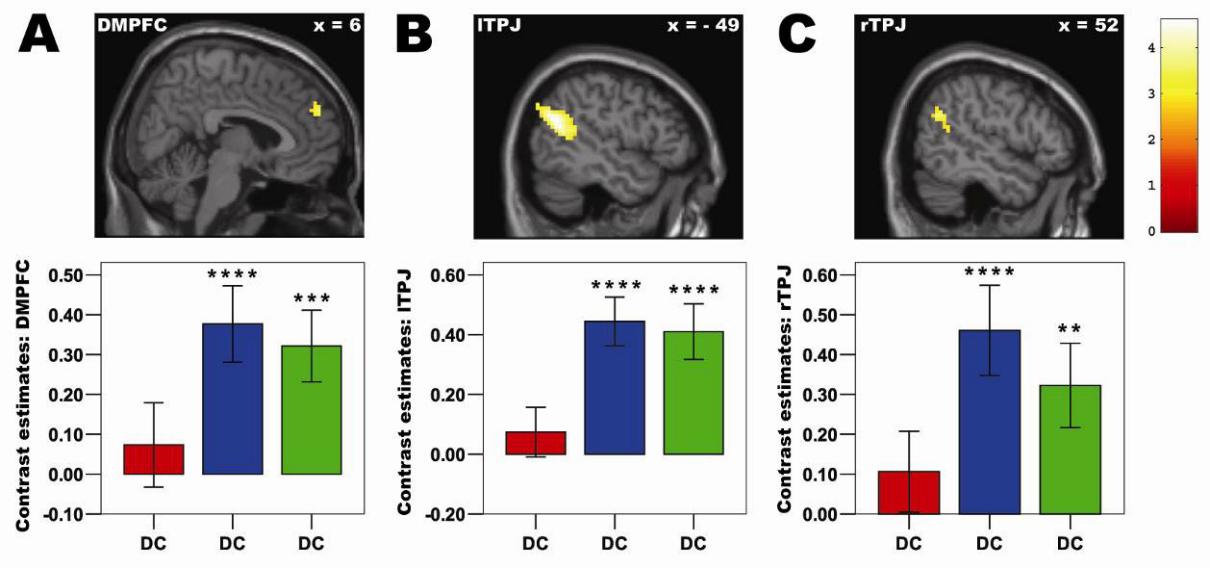

Figure 5: Ingroup effects: Mentalizing network. Depicted is the increased activation in the brain (at $p < 0.005$, voxel extent threshold: 10 voxels; activity in all regions survives small volume family-wise-error (FWE) corrections at $p < 0.05$, see methods section for details) contrasting the ingroup (AC+ABC) minus the outgroup (BC) constellations, when player A defected and player B cooperated (behavioral pattern DC). Increased activity was mainly found in brain areas involved in mentalizing processes, including **(A)** dorsomedial prefrontal cortex (DMPFC; BA 9, x = 6, y = 54, z = 30), **(B)** left temporo-parietal junction (lTPJ, BA 39/40/22, x = -45, y = -60, z = 21) and **(C)** right temporo-parietal junction (rTPJ, BA 39/40, x = 57, y = -60, z = 30). Bar plots (color coding as in Fig. 3, red = outgroup BC, blue = ingroup AC, green = ingroup ABC) representing contrast estimates (in/outgroup vs baseline) of functional or spherical ROIs (see method section for details) revealed in accordance with the similar punishment pattern that the two ingroup constellations (AC, ABC) show a highly similar activity pattern ($p > 0.25$ for all paired t-tests between the two ingroup constellations). Asterisks denote increased activity compared to baseline at $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.005$ (***) or $p < 0.001$ (****). Please see supporting Figure S2 for event-related BOLD time courses of the depicted brain regions.

Figure 6: Connectivity analyses within the mentalizing network.

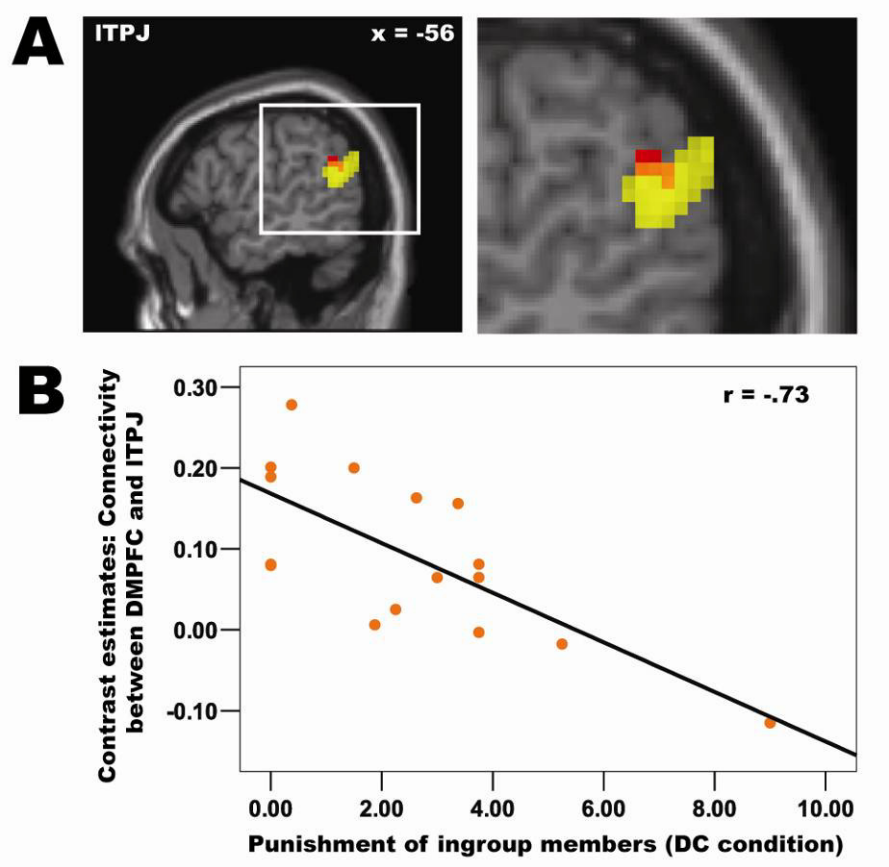

Figure 6: Connectivity analyses within the mentalizing network. (A) Connectivity analyses using the DMPFC as seed region revealed (at $p < 0.005$, voxel extent threshold = 10 voxels) that the functional connectivity between the DMPFC and lTPJ depends on third-parties' punishment level, that is the less third-parties punish defecting ingroup members in the DC condition, the stronger is the functional connectivity between these two regions. This finding provides additional evidence that the mentalizing network is recruited in order to reduce the punishment of defecting ingroup members. Color coding: connectivity effect depicted in red, activation level effect depicted in yellow (the same activation as depicted in Fig. 5B), overlap depicted in orange. **(B)** The scatter plot visualizes the effect explained in (A) using a functional ROI of the lTPJ (x = -57, y = -54, z = 24).

Figure 7: Connectivity analyses between the mentalizing-network and punishment-network.

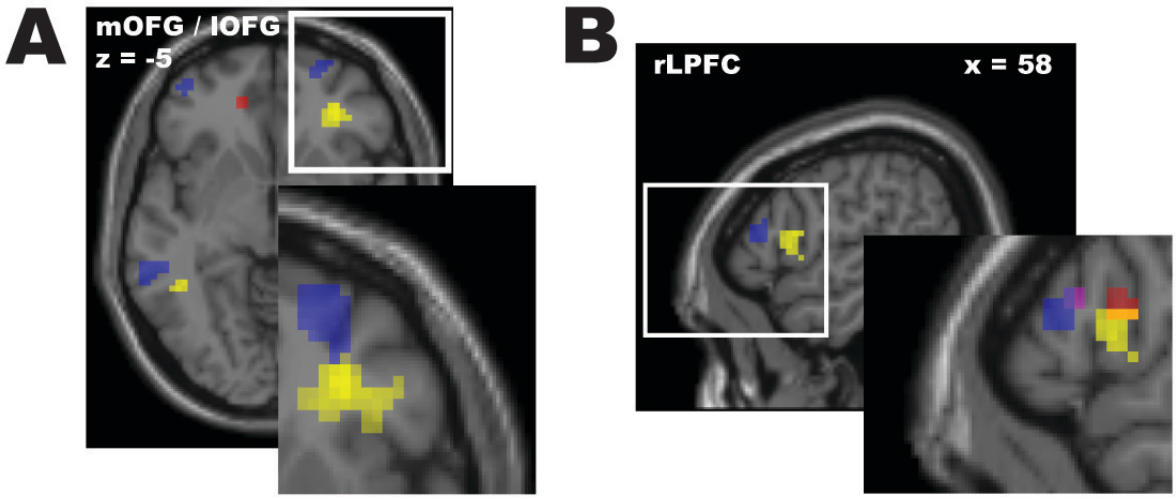

Figure 7: Connectivity analyses between the mentalizing-network and punishment-network. We applied Physio-Physiological Interaction (PPI) analyses using the DMPFC and lTPJ as seed regions in order to reveal the effective connectivity between the two networks shown to orchestrate the parochial nature of altruistic norm enforcement. Findings revealed evidence (at $p < 0.005$, voxel extent threshold: 10 voxels) that the lTPJ modulates the effective connectivity between the DMPFC and **(A)** the evaluation system in the lateral (BA 10/11, left: x = -42, y = 54, z = -9; right: x = 24, y = 60, z = -6, depicted in blue) and medial OFG (BA 10/11, x = -15, y = 45, z = -9, depicted in red) as well as **(B)** the cognitive control system in the rLPFC (BA 45/46, x = 54, y = 36, z = 15, depicted in blue). In detail, medial areas of the OFG (depicted in red) show an enhanced positive connectivity with the DMPFC whenever the lTPJ is strongly activated. In sharp contrast, lateral areas of the OFG and the rLPFC (depicted in blue) show an enhanced negative connectivity with the DMPFC whenever the lTPJ is strongly activated. These highly distinctive connectivity effects in medial and lateral areas of the OFG support our hypothesis that a justification process in the mentalizing network might change the evaluation of ingroup members' defective behavior (making it less negative and/or more positive). Notably, the negative connectivity effects are localized in neighboring and overlapping areas of the punishment-network depicted in Fig. 3 and 4. In order to visualize this spatial proximity, the same activity (depicted in yellow) and positive connectivity patterns (depicted in red and violet in the zoom view of Fig. 7B) are shown here. We in particularly want to point out (see zoom view in Fig. 7B) that the negative connectivity effect in the rLPFC (in blue) is localized in the same area showing an enhanced positive connectivity with the rOFG, whenever third-parties strongly punish defecting outgroup members (depicted in violet, overlap depicted in dark violet). For display purposes all activation and connectivity patterns in (A) and (B) are depicted at $p < 0.01$, except for the zoom view in (A) on which the patterns are depicted at $p < 0.05$.

Figure 8: Summary.

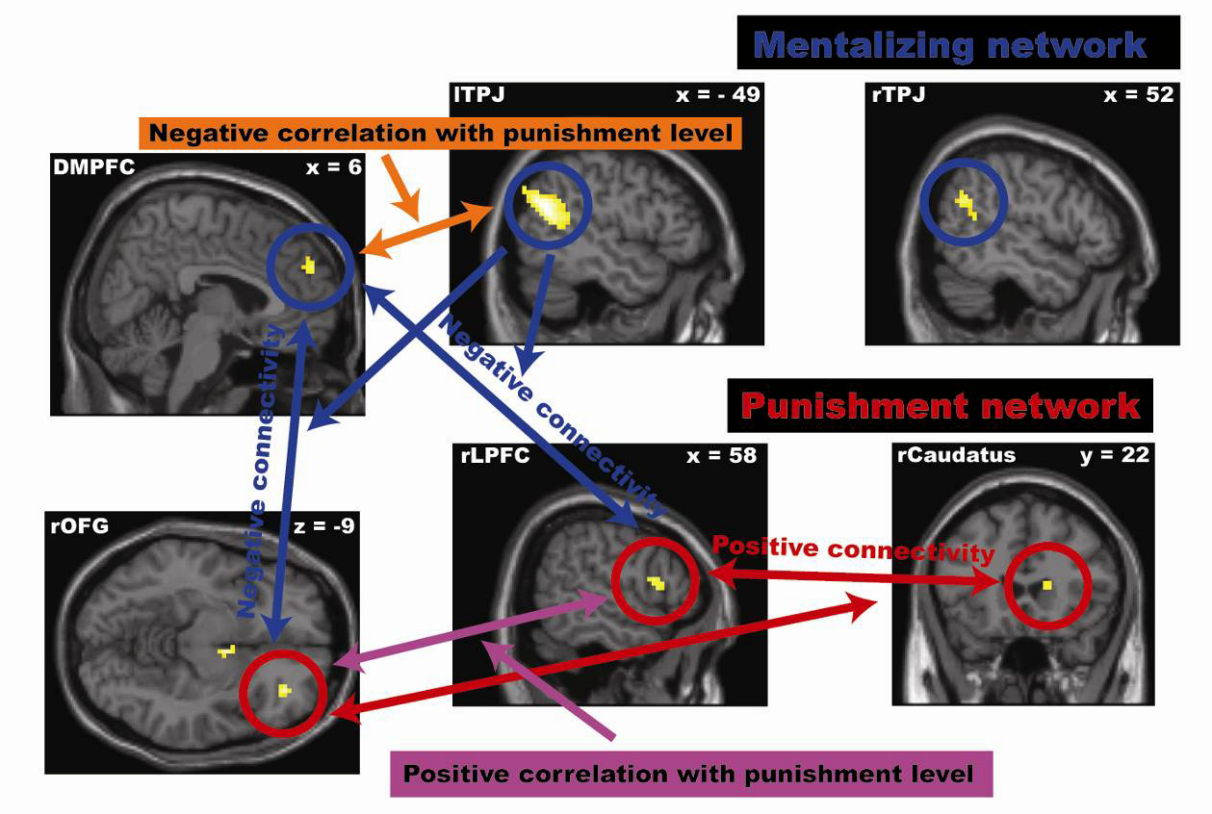

Figure 8: Summary. The analysis of the neural underpinnings of the parochial nature of altruistic norm enforcement revealed the following activity and connectivity pattern. *First,* the increased punishment of defecting outgroup members is associated with increased activity in a functionally connected network of brain areas involved in punishment-related decision processes (red circles and red lines with arrows). *Second,* the stronger the connectivity within areas of this punishment network, the stronger defecting outgroup members are punished (violet lines with arrows). *Third,* the reduced punishment of ingroup members' defective behavior is associated with increased activity in the mentalizing network of the brain, suggesting that third-parties try to understand and justify ingroup members' defective behavior (blue circles). *Fourth,* the stronger the connectivity within areas of this mentalizing network, the less third-parties punish defecting ingroup members (orange lines with arrows). *Fifth,* the analysis of connectivity between the punishment and mentalizing/justification network suggests that the mentalizing/justification process reduces the punishment behavior by modulating the activity in areas of the punishment network associated with negative evaluation processes (rOFC) and the assignment of an appropriate punishment level via the weighting of economic-self-interests (rLPFC, blue lines with arrows).

Reference list

Allport GW. (1954): The nature of prejudice.: Reading, MA: Addison-Wesley.

Baumgartner T, Fischbacher U, Feierabend A, Lutz K, Fehr E. (2009): The neural circuitry of a broken promise. Neuron 64(5):756-70.

Baumgartner T, Heinrichs M, Vonlanthen A, Fischbacher U, Fehr E. (2008): Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. Neuron 58(4):639-50.

Beer JS, Stallen M, Lombardo MV, Gonsalkorale K, Cunningham WA, Sherman JW. (2008): The Quadruple Process model approach to examining the neural underpinnings of prejudice. NeuroImage 43(4):775-83.

Bendor J, Swistak P. (2001): The evolution of norms. American Journal of Sociology 106(6):1493-1545.

Bernhard H, Fischbacher U, Fehr E. (2006): Parochial altruism in humans. Nature 442(7105):912-915.

Boyd R, Gintis H, Bowles S, Richerson PJ. (2003): The evolution of altruistic punishment. Proceedings of the National Academy of Sciences of the United States of America 100(6):3531-3535.

Brewer M. (1979): In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. Psychological Bulletin 86(2):307-324.

Brewer M. (1999): The psychology of prejudice: Ingroup love or outgroup hate? Journal of Social Issues Vol 55(3):Fal 1999.

Buckholtz JW, Asplund CL, Dux PE, Zald DH, Gore JC, Jones OD, Marois R. (2008): The neural correlates of third-party punishment. Neuron 60(5):930-40.

Chen Y, Li SX. (2009): Group Identity and Social Preferences. American Economic Review 99(1):431-457.

Choi JK, Bowles S. (2007): The coevolution of parochial altruism and war. Science (New York, N.Y 318(5850):636-40.

Cunningham WA, Johnson MK, Raye CL, Chris Gatenby J, Gore JC, Banaji MR. (2004): Separable neural components in the processing of black and white faces. Psychol Sci 15(12):806-13.

de Quervain DJ, Fischbacher U, Treyer V, Schellhammer M, Schnyder U, Buck A, Fehr E. (2004): The neural basis of altruistic punishment. Science (New York, N.Y 305(5688):1254-1258.

Delgado MR, Frank RH, Phelps EA. (2005): Perceptions of moral character modulate the neural systems of reward during the trust game. Nature Neurosci. 8(11):1611-1618.

Delgado MR, Locke HM, Stenger VA, Fiez JA. (2003): Dorsal striatum responses to reward and punishment: effects of valence and magnitude manipulations. Cogn Affect Behav Neurosci 3(1):27-38.

Eberhardt JL. (2005): Imaging race. The American psychologist 60(2):181-90.

Efferson C, Lalive R, Fehr E. (2008): The coevolution of cultural groups and ingroup favoritism. Science (New York, N.Y 321(5897):1844-9.

Fehr E, Bernhard H, Rockenbach B. (2008): Egalitarianism in young children. Nature 454(7208):1079-83.

Fehr E, Camerer CF. (2007): Social neuroeconomics: the neural circuitry of social preferences. Trends Cogn Sci 11(10):419-27.

Fehr E, Fischbacher U. (2003): The nature of human altruism. Nature 425(6960):785-791.

Fehr E, Fischbacher U. (2004a): Social norms and human cooperation. Trends in Cognitive Sciences 8(4):185-190.

Fehr E, Fischbacher U. (2004b): Third-party punishment and social norms. Evolution and Human Behavior 25(2):63-87.

Fehr E, Gächter S. (2002): Altruistic punishment in humans. Nature 415:137-140.

Fischbacher U. (2007): z-Tree: Zurich toolbox for ready-made economic experiments. Experimental Economics 10(2):171-178.

Fliessbach K, Weber B, Trautner P, Dohmen T, Sunde U, Elger CE, Falk A. (2007): Social comparison affects reward-related brain activity in the human ventral striatum. Science (New York, N.Y 318(5854):1305-8.

Frackowiak R, Friston K, Frith C, Dolan R, Price CJ, Zeki S, Ashburner J, Penny W. (2004): Human Brain Function. New York: Academic Press.

Freeman JB, Schiller D, Rule NO, Ambady N. The neural origins of superficial and individuated judgments about ingroup and outgroup members. Hum Brain Mapp 31(1):150-9.

Friston KJ, Buechel C, Fink GR, Morris J, Rolls E, Dolan RJ. (1997): Psychophysiological and modulatory interactions in neuroimaging. NeuroImage 6(3):218-29.

Gallagher HL, Frith CD. (2003): Functional imaging of 'theory of mind'. Trends Cogn Sci 7(2):77-83.

Goette L, Huffman D, Meier S. (2006): The Impact of Group Membership on Cooperation and Norm Enforcement: Evidence Using Random Assignment to Real Social Groups. American Economic Review 96 2:212-16.

Golby AJ, Gabrieli JD, Chiao JY, Eberhardt JL. (2001): Differential responses in the fusiform region to same-race and other-race faces. Nature neuroscience 4(8):845-50.

Hampton AN, Bossaerts P, O'Doherty JP. (2008): Neural correlates of mentalizing-related computations during strategic interactions in humans. Proceedings of the National Academy of Sciences of the United States of America 105(18):6741-6.

Harbaugh WT, Mayr U, Burghart DR. (2007): Neural Responses to Taxation and Voluntry Giving Reveal Motives for Charitable Donations. Science (New York, N.Y 316:1622-1625.

Hare TA, Camerer CF, Rangel A. (2009): Self-control in decision-making involves modulation of the vmPFC valuation system. Science (New York, N.Y 324(5927):646-8.

Harris LT, Fiske ST. (2006): Dehumanizing the lowest of the low: neuroimaging responses to extreme out-groups. Psychol Sci 17(10):847-53.

Haslam N. (2006): Dehumanization: An integrative review. Personality and Social Psychology Review 10(3):252-264.

Henrich J. (2006): Cooperation, punishment, and the evolution of human institutions. Science (New York, N.Y 312(5770):60-61.

Henrich J, McElreath R, Barr A, Ensminger J, Barrett C, Bolyanatz A, Cardenas JC, Gurven M, Gwako E, Henrich N, Lesorogol C, Marlowe F, Tracer D, Ziker J. (2006): Costly punishment across human societies. Science (New York, N.Y 312(5781):1767-1770.

Hewstone M, Rubin M, Willis H. (2002): Intergroup bias. Annual Review of Psychology 53:575-604.

Hill K. (2002): Altruistic Cooperation During Foraging by the Ache, and the evolved Human Predisposition to Cooperate. Human Nature 13(1):105-128.

Holm S. (1979): A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistic 6:65-70.

Ito TA, Bartholow BD. (2009): The neural correlates of race. Trends Cogn Sci 13(12):524-31.

Kaplan H, Hill J, Lancaster J, Hurtado AM. (2000): A theory of human life history evolution: Diet, intelligence, and longevity. Evol Anthropol 9(4):156-185.

Kawagoe R, Takikawa Y, Hikosaka O. (1998): Expectation of reward modulates cognitive signals in the basal ganglia. Nature neuroscience 1(5):411-416.

King-Casas B, Tomlin D, Anen C, Camerer CF, Quartz SR, Montague PR. (2005): Getting to Know You: Reputation and Trust in a Two-Person Economic Exchange. Science (New York, N.Y 308(5718):78-83.

Kinzler KD, Dupoux E, Spelke ES. (2007): The native language of social cognition. Proceedings of the National Academy of Sciences of the United States of America 104(30):12577-12580.

Knoch D, Nitsche MA, Fischbacher U, Eisenegger C, Pascual-Leone A, Fehr E. (2007): Studying the Neurobiology of Social Interaction with Transcranial Direct Current Stimulation The Example of Punishing Unfairness. Cereb Cortex Advance Access(December 24, 2007).

Knoch D, Pascual-Leone A, Meyer K, Treyer V, Fehr E. (2006): Diminishing reciprocal fairness by disrupting the right prefrontal cortex. Science (New York, N.Y 314(5800):829-832.

Koopmans R, Rebers S. (2009): Collective action in culturally similar and dissimilar groups: an experiment on parochialism, conditional cooperation, and their linkages. Evolution and Human Behavior 30(3):201-211.

Kringelbach ML. (2005): The human orbitofrontal cortex: linking reward to hedonic experience. Nature reviews 6(9):691-702.

Lieberman MD, Cunningham WA. (2009): Type I and Type II error concerns in fMRI research: re-balancing the scale. Soc Cogn Affect Neurosci 4(4):423-8.

Lieberman MD, Hariri A, Jarcho JM, Eisenberger NI, Bookheimer SY. (2005): An fMRI investigation of race-related amygdala activity in African-American and Caucasian-American individuals. Nature neuroscience 8(6):720-2.

Liu X, Powell DK, Wang H, Gold BT, Corbly CR, Joseph JE. (2007): Functional dissociation in frontal and striatal areas for processing of positive and negative reward information. J Neurosci 27(17):4587-97.

Maldjian JA, Laurienti PJ, Kraft RA, Burdette JH. (2003): An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. NeuroImage 19(3):1233-9.

Mathur VA, Harada T, Lipke T, Chiao JY. Neural basis of extraordinary empathy and altruistic motivation. NeuroImage 51(4):1468-75.

Mitchell JP, Neil Macrae C, Banaji MR. (2005): Forming impressions of people versus inanimate objects: social-cognitive processing in the medial prefrontal cortex. NeuroImage 26(1):251-7.

O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ. (2004): Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science (New York, N.Y 304(5669):452-454.

Phelps EA, O'Connor KJ, Cunningham WA, Funayama ES, Gatenby JC, Gore JC, Banaji MR. (2000): Performance on indirect measures of race evaluation predicts amygdala activation. Journal of cognitive neuroscience 12(5):729-38.

Plassmann H, O'Doherty J, Rangel A. (2007): Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. J Neurosci 27(37):9984-8.

Pruessmann KP, Weiger M, Scheidegger MB, Boesiger P. (1999): SENSE: sensitivity encoding for fast MRI. Magn Reson Med 42(5):952-62.

Rangel A, Camerer C, Montague PR. (2008): A framework for studying the neurobiology of value-based decision making. Nature reviews 9(7):545-56.

Richeson JA, Baird AA, Gordon HL, Heatherton TF, Wyland CL, Trawalter S, Shelton JN. (2003): An fMRI investigation of the impact of interracial contact on executive function. Nature neuroscience 6(12):1323-8.

Rilling J, Gutman D, Zeh T, Pagnoni G, Berns G, Kilts C. (2002): A neural basis for social cooperation. Neuron 35(2):395-405.

Rilling JK, Goldsmith DR, Glenn AL, Jairam MR, Elfenbein HA, Dagenais JE, Murdock CD, Pagnoni G. (2007): The neural correlates of the affective response to unreciprocated cooperation. Neuropsychologia.

Rilling JK, Sanfey AG, Aronson JA, Nystrom LE, Cohen JD. (2004): The neural correlates of theory of mind within interpersonal interactions. NeuroImage 22(4):1694-703.

Sanfey AG. (2007): Social decision-making: insights from game theory and neuroscience. Science (New York, N.Y 318(5850):598-602.

Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD. (2003): The neural basis of economic decision-making in the Ultimatum Game. Science (New York, N.Y 300(5626):1755-1758.

Schultz W, Romo R. (1988): Neuronal-Activity in the Monkey Striatum During the Initiation of Movements. Exp Brain Res 71(2):431-436.

Seymour B, Singer T, Dolan R. (2007): The neurobiology of punishment. Nature reviews 8(4):300-11.

Singer T, Seymour B, O'Doherty JP, Stephan KE, Dolan RJ, Frith CD. (2006): Empathic neural responses are modulated by the perceived fairness of others. Nature 439(7075):466-469.

Sober E, Wilson DS. (1998): Unto Others - The Evolution and Psychology of Unselfish Behavior. Cambridge, Massachusetts: Harvard University Press. 394 p.

Spitzer M, Fischbacher U, Herrnberger B, Gron G, Fehr E. (2007): The neural signature of social norm compliance. Neuron 56(1):185-96.

Steinbeis N, Koelsch S. (2009): Understanding the intentions behind man-made products elicits neural activity in areas dedicated to mental state attribution. Cereb Cortex 19(3):619-23.

Strobel A, Zimmermann J, Schmitz A, Reuter M, Lis S, Windmann S, Kirsch P. (2011): Beyond revenge: neural and genetic bases of altruistic punishment. NeuroImage 54(1):671-80.

Tajfel H, Billig M, Bundy R, Flament C. (1971): Social Categorization in Intergroup Behaviour. European Journal of Social Psychology 1:149-178.

Tajfel H, Turner JC. (1979): An Integrative Theory of Intergroup Conflict. In: Austin WG, Worchel S, editors. The Psychology of Intergroup Relations. Monterey: Nelson Hall.

van 't Wout M, Kahn RS, Sanfey AG, Aleman A. (2005): Repetitive transcranial magnetic stimulation over the right dorsolateral prefrontal cortex affects strategic decision-making. Neuroreport 16(16):1849-52.

Van Bavel JJ, Packer DJ, Cunningham WA. (2008): The neural substrates of in-group bias: a functional magnetic resonance imaging investigation. Psychol Sci 19(11):1131-9.

Van Overwalle F. (2009): Social cognition and the brain: a meta-analysis. Hum Brain Mapp 30(3):829-58.